Retrieval-Augmented Fine-Tuning Bypasses the Parametric Knowledge Ceiling: Benchmarking Zero-Hallucination Machining-Copilot LLMs on 8×H100

U2DIA AI Research (Suyong Yun)

U2DIA AI, Incheon, Republic of Korea · syyun@u2dia-ai.com

Technical Report TR-2026-01 · Program: NVIDIA Innovation Lab, 2026-05-12 → 2026-07-11 (60 days) · Report date 2026-07-13

ABSTRACT

Large language models deployed beside running CNC machines must satisfy a stricter contract than general assistants: an unsupported feed/speed, alarm, or canned-cycle answer can cause physical damage, so a false answer is strictly worse than a refusal. We study whether an open-weight LLM can act as a *zero-hallucination, sub-3-second machining copilot* for NC-code, cutting-condition and controller questions across seven controller dialects. Over a 60-day program on an 8×H100 SXM5 cluster we fine-tune Qwen-family models at 8B, 32B and 72B scales and evaluate every model × serving configuration to the individual-run level. Our central finding is diagnostic: the residual hallucination that closed-book adapters could not remove is a *knowledge* problem, not a *behavior* problem — 457 closed-book LoRA adapters failed to lift trap-set correctness beyond a 4–10/100 parametric ceiling. Retrieval-Augmented Fine-Tuning (RAFT) over a 220,813-row cited corpus bypasses this ceiling at every scale: the identical 8B base improves from 59.7 correct / 90.5% false-answer (SFT) to 75.2 / 9.5% (RAFT) at unchanged cost, and a 72B tier reaches 88.4–89.9 correct with 100% source-grounding and 4.8% false-answer. Offline per-channel FP8 (W8A8) quantization places the 72B tier on a real-time envelope (p99 1.73 s, ~18 GiB/GPU, ≈13× serving-cost reduction) with no measurable quality loss, whereas per-tensor dynamic FP8 collapses to degenerate output. On a strictly held-out evaluation (n=200 unseen alarm distribution), false answers fall to 1.5% with 98.5% correct refusal — evidence of learned *read → ground → cite, else abstain* behavior rather than memorization. Because in-weights refusal saturates near 60% regardless of no-answer training share, we adopt a retrieval-confidence gate in front of generation, driving effective production hallucination toward zero. We release our measurement methodology, including a failure mode in which closed-book harnesses reward fabricated citations and mis-score RAFT models.

Keywords: retrieval-augmented fine-tuning · hallucination · refusal · direct preference optimization · LoRA · FP8 quantization · vLLM · CNC machining · domain-specific LLM

1 Introduction

General-purpose LLMs answer manufacturing questions fluently but unreliably: on our adversarial trap set they fabricate G/M-codes, alarm semantics and cutting parameters with high confidence. In office software a wrong answer costs a correction; on a

shop floor it can cost a spindle. The deployment contract for a machining copilot is therefore asymmetric: *the model may only speak when a verifiable source supports the answer, must cite that source, and must abstain otherwise* — all within an interactive latency budget.

This report asks three questions. **(Q1)** Can domain fine-tuning alone (closed-book SFT / preference optimization) reach production-grade correctness on hostile, dialect-specific machining questions? **(Q2)** If not, does Retrieval-Augmented Fine-Tuning (RAFT) — training the model to answer *over supplied documents* — change the picture, and does the effect hold from 8B to 72B? **(Q3)** Can the resulting quality be served in real time on commodity multi-GPU nodes, and at what cost per token?

Our answers, measured rather than argued: (Q1) no — a 4–10/100 parametric ceiling persists across 457 closed-book adapters; (Q2) yes — RAFT lifts correctness by an order of magnitude at every scale and cuts the 8B false-answer rate by 81 points at unchanged serving cost; (Q3) yes — offline per-channel FP8 collapses 72B serving cost from 13.0× to ≈1.0× of the 32B reference while *raising* measured correctness.

2 Related Work

Retrieval-augmented generation grounds generation in retrieved evidence [1]; RAFT [2] moves the grounding skill into the weights by fine-tuning with golden and distractor documents in-context. Direct Preference Optimization [3] provides a reward-model-free preference-learning objective which we use for our second-stage training. Low-Rank Adaptation [4] makes 72B-scale fine-tuning tractable on a single 8-GPU node. Our serving stack builds on vLLM/PagedAttention [5]; our quantization study follows the FP8 formats introduced in [6] and is implemented with offline per-channel W8A8 calibration. Base models are Qwen2.5-72B-Instruct [7] and Qwen3-8B/32B [8]. We do not claim novelty for any individual component; the contribution is a controlled, atomic benchmark of the *combination* under a safety-critical domain contract, plus two negative results (the parametric ceiling; per-tensor dynamic FP8 collapse) and a measurement-methodology correction (§6.3).

3 Task, Corpora and Knowledge Base

3.1 Task definition

The task is question answering over CNC machining knowledge — NC-code semantics, canned cycles, cutting conditions, alarms and controller-specific behavior — spanning seven controller dialects (Fanuc, Siemens Sinumerik, Heidenhain, Mazak, Okuma, Haas, Brother). Given a question and (in the RAG setting) a retrieved document, the model must produce an answer supported by the document with an explicit `[source: …]` citation, or state that the information is insufficient. Target latency is an interactive envelope with p99 ≤ 4.5 s under load.

3.2 Training corpora

- **RAFT corpus v3** (`u2dia_nc_v3_raft.jsonl`): 220,813 rows in messages format with citations; each row is a golden-document, distractor-document, or insufficient-information case, so the model learns grounding, distractor rejection and abstention jointly.
- **Synthetic hallucination-trap corpus**: 100K+ generated questions with 29.7% deliberate no-answer cases; integrity-verified with zero leakage into evaluation sets.
- **Controller-spec knowledge base**: 7 controllers, 345 grounded facts (34 caveats) — the retrieval substrate used both in production and in the RAG-context evaluation harness.

3.3 Evaluation sets

(i) An adversarial **trap set** targeting fabrication-prone semantics (look-alike codes, dialect divergence); (ii) a **production-configuration RAG set** mirroring the serving pipeline; (iii) a strictly **held-out set** (n=200) drawn from an unseen alarm distribution whose documents never appear in training (§7.4).

4 Models and Training

4.1 Hardware and orchestration

All training ran on a single node with 8×H100 SXM5 80GB (NVLink) provisioned through NVIDIA Brev (instance `top-magenta-wolf`, org `nvidia-u2dia`). GPUs 0–6 executed our workloads; GPU 7 was reserved for an unrelated concurrent track and never touched by our dispatcher. A self-healing dispatcher kept the fleet saturated (automatic gap-fill on idle GPUs, non-leader checkpoint archiving under disk pressure, and a four-way launch safety net: wall-clock, cost, ticket-driven, dead-man).

4.2 Training configurations

All production models are LoRA adapters [4] (PEFT 0.19.1, task `CAUSAL_LM`, dropout 0.05, bias none) over open-weight bases, trained with TRL: a supervised RAFT stage (SFT) followed by preference optimization (DPO [3]). Table 1 lists the four leader runs whose artifacts were archived at program close.

Leader run	Base model	Stage	LoRA r	α	Target modules	Adapter size	Date (UTC)
`manual-sft-72b-r4`	Qwen2.5-72B-Instruct	SFT (TRL)	80	160	`q,k,v,o,gate,up,down proj`	2.11 GB	2026-06-21
`manual-rl-72b-r4 (run 1)`	Qwen2.5-72B-Instruct	DPO	80	160	`same 7 proj`	2.11 GB	2026-06-21
`manual-rl-72b-r4 (run 2)`	Qwen2.5-72B-Instruct	DPO	80	160	`same 7 proj`	2.11 GB	2026-06-26
`manual-rl-8b-r25`	Qwen3-8B	DPO	64	128	`same 7 proj`	333 MB	2026-06-26

Table 1. Archived leader runs. All adapters target the seven projection matrices (attention q/k/v/o + MLP gate/up/down). The 32B production model (`raft32b`, Qwen3-32B, LoRA r=32/α=64) is retained in the serving registry. RAFT ceiling on 72B was reached at 80K samples / 2 epochs; additional data or epochs over-fit back to 85–87 correct.

4.3 Training telemetry (exploratory loop)

Alongside the production models we ran an exploratory continued-pretraining loop at sequence length 8192, LoRA rank 32, LR 2e-4. Best clean convergence: run `v61` (min loss 0.226 over 424 steps / 15.7 h); longest stable run: `v74` (final loss 0.739, 20.04 h). Two MoE-scale runs (`v63`/`v64`) failed with CUDA OOM at step 1 of 1341 — attempting a 3.59 GiB allocation with ~420 MiB free of 79.18 GiB — and were recovered by kernel fallbacks and sequence-length reduction (8192 → 1024 for the pathological configurations). At the 80 GB boundary, kernel fallback paths and sequence budgeting proved mandatory rather than optional.

5 Serving and Quantization

5.1 Serving engine

All reported serving numbers use vLLM [5]: 8B on TP1 (`enforce_eager`), 32B on TP2 with CUDA-graph capture, 72B on TP4 (bf16 and FP8 variants). LoRA adapters are merged for the persistent FP8 endpoint (cold load ≈122 s). FP8 serving requires `VLLM_USE_DEEP_GEMM=0` and `VLLM_DEEP_GEMM_WARMUP=skip` in our environment.

5.2 Quantization study

The 72B tier is interactive only when quantized, but the quantization *method* dominates the outcome (Table 2): offline per-channel W8A8 calibration (llm-compressor, isolated virtualenv due to a dependency conflict with vLLM) preserves quality within greedy-decode noise at roughly half the memory, while per-tensor *dynamic* FP8 — attractive on latency — produces degenerate output because a single per-tensor scale cannot track activation outliers.

72B variant	Method	p99 latency	Mem / GPU	Correct	Outcome
bf16 (reference)	full precision, TP4	~21 s	~38 GiB	88.4	accuracy reference; non-interactive
dynamic FP8	per-tensor, on-the-fly	~2.37 s	35.6 GiB	degenerate	rejected — scale cannot track outliers
offline FP8 (W8A8)	per-channel, calibrated	**1.73 s (TP4) / ~3.8 s (TP2)**	**~18 GiB**	**89.9**	adopted — within noise of bf16

Table 2. Quantization ablation on the 72B RAFT model. Per-channel offline calibration is the difference between collapse and a real-time maximum-quality tier; correctness 89.9 vs the bf16 88.4 reference lies inside greedy-decode noise.

6 Evaluation Methodology

6.1 Metrics

- **Correct** — answer is supported by the supplied source document (0–100).
- **Grounding** — the answer cites the supplied source.
- **Refusal accuracy** — the model correctly abstains when the source does not contain the answer.
- **False-answer (hallucination)** — an unsupported assertion; the safety-critical metric.
- **Latency** — full-response wall-clock; p50 at a representative ~52 output tokens, p99/worst at the longest observed ~220-token response on the same engine.
- **Relative cost** — GPU-seconds per output token, (TP GPU count)/(decode tok/s), normalized to the 32B TP2 real-time configuration = 1.00×. A serving-cost proxy; the cluster itself was grant-funded.

6.2 RAG-context harness

The production-faithful harness (`raft_rag_eval.py`) supplies the retrieved document in-prompt — exactly as served — and scores the four quality metrics above. A scoring defect was found and fixed during the program: correct refusals phrased as "insufficient information / no source" were being counted as fabricated citations until a non-source token filter (`none` / `n/a` / `no source`) corrected refusal accounting.

6.3 Why closed-book harnesses mis-score RAFT models

> **Methodology finding.** A closed-book harness presents trap questions with *no* document. A non-RAFT SFT model emits a `[source: …]` citation *by habit* even when no source exists — scoring ~98–100% "grounding" that is entirely fabricated. A RAFT model honestly emits no citation when no document is given, and is *penalized* for it. Closed-book evaluation therefore rewards citation mimicry and punishes true grounding; all RAFT results in this report use the RAG-context harness. We regard the harness itself as a first-class research artifact: bad metrics hide good models.

7 Results

7.1 Master benchmark matrix

Model / config	Base	p50	p99/worst	tok/s	GPUs (TP)	Correct	Ground	Refusal	False-ans ↓	Rel. cost
8B SFT · TP1 eager	Qwen3-8B	~1.9 s	—	81–84	1	59.7	90.7	—	90.5	0.29×
8B RAFT · TP1	Qwen3-8B	~1.9 s	—	81–84	1	75.2	99.2	—	9.5	**0.29×**
32B RAFT · TP2 CUDA-graph	Qwen3-32B	**1.1 s**	4.56 s	47–48	2	79.1	**100**	90.5	9.5	1.00×
72B RAFT · TP4 bf16	Qwen2.5-72B-Inst	~21 s	—	7.3	4	88.4	**100**	61.0	4.8	13.0×
72B offline-FP8 · TP4	Qwen2.5-72B-Inst	**1.73 s**	—	high	4 (or 2)	**89.9**	**100**	90.5	**4.8**	~1.0×
72B dynamic-FP8 · TP4	Qwen2.5-72B-Inst	~2.37 s	—	92.8	4	degenerate	—	—	collapse	1.02×

Table 3. Master matrix, all rows vLLM-measured on the same node. Un-measured cells (—) are left un-imputed. The 72B 4.8% false-answer figure and the 32B/8B 9.5% figures come from the same retrieval setting; the gap is the model-size correctness premium.

7.2 The hallucination curve is governed by serving regime, not adapter count

Regime	Correct	Note
Closed-book, any adapter count (457 adapters, SFT/cDPO)	4–10	parametric ceiling; deeper cDPO chains improved a combined trap/self metric 130 → 52 but trap stayed at 44%
8B SFT + retrieval (no RAFT)	59.7	retrieval alone is insufficient — 90.5% false-answer
8B RAFT + retrieval	75.2	−81-pt false-answer reduction vs SFT at equal cost/speed
32B RAFT + retrieval	79.1	production real-time configuration
72B RAFT + retrieval (bf16)	88.4	accuracy reference
72B RAFT + retrieval (offline FP8)	**89.9**	maximum-quality real-time tier

Table 4. Answer correctness (inverse of hallucination) across serving regimes. One RAFT pass over retrieval dominates any amount of closed-book adapter training.

7.3 Sustained-load serving (72B-FP8)

Concurrency	p99	TTFT	Req/s	Agg. decode tok/s	Served correct*
c = 1 (single stream)	1.73 s	–	–	91	89.9 (HF-direct)
c = 96 (recommended operating point)	4.02 s	424 ms	25	3,939	~85
c = 128 (stress)	4.16 s	–	–	–	~85 · 0 SLA failures

Table 5. Sustained load against a 4.5 s p99 SLA. (*) Served correctness ~85 (84.5–85.3) vs 89.9 single-shot HF-direct; the delta is served/FP8 cost plus greedy noise. Grounding 100 and false-answer 4.8 reproduced under load; cross-size served re-checks reproduced (32B 82.2 ≈ 79.1; 8B 74.4 ≈ 75.2). Decode dominates wall-clock: capping answers at ≈90 output tokens is the strongest single SLA lever, ahead of TP topology.

7.4 Held-out generalization

A standing concern was training-set overlap inflating in-distribution scores. On the strictly held-out set (n=200, unseen alarm distribution, documents never seen in training) the 32B RAFT model produces **1.5% false answers** and **98.5% correct refusal** — *better* than in-distribution. A memorizing model would confabulate on unseen documents; the observed behavior is the signature of a learned read–ground–cite–abstain policy. An apparent in-distribution "correct" drop under out-of-distribution answer shapes was traced by manual audit to citation-token/paraphrase measurement artifacts (zero genuinely wrong answers) and eliminated by pinning the citation template in the system prompt.

7.5 Refusal: in-weights ceiling and the retrieval-confidence gate

Lever	Observation	Consequence
No-answer training share 11.5% → 19% → 35%	both refusal *and* correctness declined as the share rose	over-weighting abstention rows is counter-productive
In-weights refusal	saturates ≈60% (72B: 61.0), scale-independent	weights alone cannot reach production refusal
Retrieval-confidence gate (adopted)	98.5% held-out correct refusal with the gate	abstain *before* generation when support is weak

Table 6. Refusal is architecturally gated at retrieval, not trained into the weights. In production the model is only invoked when a supporting document clears the confidence threshold, and it must cite that document — driving effective hallucination toward zero without fighting the intrinsic ceiling.

7.6 Cost model

Tier	GPUs	tok/s	Rel. cost	Interpretation
8B RAFT (budget)	1	81–84	0.29×	highest tok/s per GPU; on-device candidate
32B RAFT (real-time)	2	47–48	1.00×	reference; latency-optimal at ≤ ~90-token answers
72B offline-FP8 (max quality)	4 (or 2)	high	~1.0×	72B accuracy at ≈32B cost — the economic unlock
72B bf16 (batch)	4	7.3	13.0×	accuracy ceiling; non-interactive

Table 7. GPU-seconds per output token, normalized. Offline FP8 collapses 72B serving cost 13.0× → ≈1.0× while raising measured correctness (88.4 → 89.9): the top-accuracy and reference-cost tiers converge.

8 Analysis and Discussion

- **Retrieval > parameters.** Trustworthiness is dominated by grounding-on-retrieval, not model size: an 8B RAFT model out-refuses larger closed-book models.
- **RAFT is size-agnostic.** The ceiling bypass reproduces at 8B, 32B and 72B, de-risking a single-GPU budget tier below 10% false-answer.

- **Quantization method > bit-width.** Per-channel offline FP8 is the difference between collapse at 2.4 s and 89.9 correct at 1.73 s.
- **Output length is the SLA lever.** Serving is decode-bound; a ~90-token cap (or first-token streaming) buys more latency headroom than any TP change.
- **Safety generalizes out-of-distribution.** Held-out false-answer (1.5%) is *lower* than in-distribution — the abstention policy transferred.
- **The harness is part of the system.** A mis-specified metric inverted the ranking of SFT vs RAFT models (§6.3); measurement tooling deserves the same rigor as training.

9 Limitations

- **R&D scope.** All measurements are research benchmarks on the grant cluster; no production traffic ran on it, and production serving operates on separate infrastructure. This report is a sizing/adoption study, not a deployment record.
- **Regime comparability.** RAFT-RAG figures and closed-book curriculum figures are distinct measurement regimes and are not directly comparable; we report both without imputation.
- **Un-measured cells.** TensorRT-LLM AWQ engines were not built (no `nvcc` on the provisioned image; deferred to an NGC container), and the 72B-FP8 TP2 full throughput sweep is pending (p99 ≈3.8 s confirmed). We report these as open items rather than substituting numbers.
- **Greedy-decode noise.** Single-shot correctness differences of ±1–2 points are within noise; served-vs-direct deltas (~85 vs 89.9) include serving-stack cost.
- **Domain breadth.** Seven controller dialects and one language domain; transfer to other manufacturing verticals is future work.

10 Conclusion

Within a 60-day window on one 8×H100 node, retrieval-augmented fine-tuning turned open-weight Qwen models into machining copilots that answer only when evidence supports them: 89.9/100 grounded correctness with 100% citation grounding and 4.8% false answers at a 1.73 s p99 (72B, offline FP8), 1.5% false answers on strictly held-out documents, and a retrieval-confidence gate that supplies the refusal reliability the weights intrinsically lack. The parametric ceiling result (4–10/100 across 457 closed-book adapters) suggests that for safety-critical industrial domains, investment in the knowledge base and retrieval gate dominates investment in ever-larger closed-book fine-tunes.

A Reproducibility Details

- **Cluster:** 8×H100 SXM5 80GB NVLink, single node (NVIDIA Brev, `top-magenta-wolf`). GPUs 0–6 ours; GPU 7 out of scope.
- **Frameworks:** TRL (SFT/DPO) · PEFT 0.19.1 (LoRA) · vLLM serving · llm-compressor (offline W8A8, isolated virtualenv due to dependency conflict with vLLM).
- **Serving env:** `VLLM_USE_DEEP_GEMM=0`, `VLLM_DEEP_GEMM_WARMUP=skip`; 8B `enforce_eager` TP1; 32B CUDA-graph TP2; 72B TP4. Persistent FP8 endpoint cold-load ≈122 s.
- **Eval env:** `HF_HUB_OFFLINE=1`, isolated datasets cache, unbuffered execution; RAG-context harness `raft_rag_eval.py` with the non-source token filter (`none/n/a/no source`).
- **Latency protocol:** full-response wall-clock; p50 at ~52 output tokens, p99/worst at ~220 output tokens, same engine.
- **Cost protocol:** (TP GPU count)/(decode tok/s) per output token, normalized to 32B TP2 = 1.00×.

B Released / Archived Artifacts

- **Model artifacts (archived, verified object-side):** four leader runs — adapter + config + tokenizer, 25 objects / 6.25 GB — off-boarded to durable cloud object storage at program close (node teardown 2026-07-13), with a production-loop manifest (2026-07-07). Compact adapters additionally mirrored to a private model hub; the 71 GB merged FP8 checkpoint is self-hosted.
- **Corpora:** RAFT v3 (220,813 rows, cited; golden/distractor/insufficient) · 100K+ synthetic trap corpus (29.7% no-answer, leak-checked) · controller KB (7 dialects, 345 grounded facts).
- **Planned community release:** the hallucination-trap evaluation harness (golden trap set + 200 traps) and the RAG-context evaluator (`raft_rag_eval.py`).
- **Access:** artifacts and benchmark logs are available for partner technical review under NDA.

References

1. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *NeurIPS*, 2020.
2. T. Zhang, S. G. Patil et al., "RAFT: Adapting Language Model to Domain Specific RAG," arXiv:2403.10131, 2024.
3. R. Rafailov et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," *NeurIPS*, 2023 (arXiv:2305.18290).
4. E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685, 2021.
5. W. Kwon et al., "Efficient Memory Management for Large Language Model Serving with PagedAttention," *SOSP*, 2023.
6. P. Micikevicius et al., "FP8 Formats for Deep Learning," arXiv:2209.05433, 2022.
7. Qwen Team, "Qwen2.5 Technical Report," arXiv:2412.15115, 2024.
8. Qwen Team, "Qwen3 Technical Report," arXiv:2505.09388, 2025.

DION AI Technical Report TR-2026-01 · NVIDIA Innovation Lab program 2026-05-12 → 2026-07-11 · All figures drawn from the program's internal benchmark logs, serving-benchmark record and production-model decision record. R&D artifacts only; no production traffic ran on the lab environment.

Contact: syyun@u2dia-ai.com